Emotional Prosody Measurement (EPM): A Voice-Based Evaluation Method for Psychological Therapy Effectiveness

Egon L. VAN DEN BROEK
Nijmegen Institute for Cognition and Information, University of Nijmegen
P.O. Box 9104, 6500 HE Nijmegen, The Netherlands

Abstract. The voice embodies three sources of information: speech, the identity, and the emotional state of the speaker (i.e., emotional prosody). The latter feature is resembled by the variability of the F0 (also named fundamental frequency of pitch) (SD F0). To extract this feature, Emotional Prosody Measurement (EPM) was developed, which consists of 1) speech recording, 2) removal of speckle noise, 3) a Fourier Transform to extract the F0-signal, and 4) the determination of SD F0. After a pilot study in which six participants mimicked emotions by their voice, the core experiment was conducted to see whether EPM is successful. Twenty-five patients suffering from a panic disorder with agoraphobia participated. Two methods (story-telling and reliving) were used to trigger anxiety and were compared with comparable but more relaxed conditions. This resulted in a unique database of speech samples that was used to compare the EPM with the Subjective Unit of Distress to validate it as measure for anxiety/stress. The experimental manipulation of anxiety proved to be successful and EPM proved to be a successful evaluation method for psychological therapy effectiveness.

1. Introduction

Already more than a century ago Helmholtz [1] noted that one's state of mind is mirrored by characteristics of one's voice. Since half a century [2,3,4] the research towards expression of emotions increased. In 1956 Brunswik [5] introduced his lens model of perception, which is modified to a model of vocal communication of emotion [6,7]. Nowadays we know that minute inter- and intra-individual variations of the generic structure of speech indeed carry useful information [8]. This empowers us to recognize individuals and emotional states by voice [9].

The rapid variations in various acoustic parameters of human speech enable us to provide additional information next to speech itself [4]. The source of all this information is exhaled air from the lungs that provides power to drive oscillations of the vocal folds (or 'vocal cords'), which are located in the larynx.

The rate of vocal fold oscillation (respectively, about 100 Hz in adult men and about 200 Hz in adult women) determines the F0 (also named fundamental frequency of pitch) of the sound thus produced. The acoustic energy generated then passes through the vocal tract (the pharyngeal, oral and nasal cavities), where it is filtered, and finally gets out to the environment through the nostrils and lips. This filtering process plays a crucial role in speech. The filtering is accomplished by a series of bandpass filters, which are termed formants. The formants modify the sound that is emitted, allow specific frequencies to pass unhindered, and block the transmission of others. Formants are determined by the length

and shape of the vocal tract, and are rapidly modified during speech by moving the articulators (tongue, lips, soft palate, etc.).

It is important to note that these formants are independent of the F0. The F0 is determined by the vibration rate of the vocal folds (the source). Formants, on the other hand, are determined by the vocal tract (the filter). The independence of source and filter is one of the key insights of modern speech acoustics [10].

The present study focuses on F0, which carries the affective information of one's voice, i.e., the *emotional prosody* [11]. Prosody is the "rhythmic and intonational aspect of language"[1]. Emotional prosody is a set of acoustic parameters of speech directly influenced by affect such as mean amplitude, segment and pause duration, mean F0, and F0 variation (see [12] for a review on acoustic parameters). It allows the listener to infer much of the speaker's affective state [12]. In literature F0 is most often mentioned as the most successful measure for the determination of emotional prosody. We have, therefore, chosen to use F0 for *Emotional Prosody Measurement (EPM)* on behalf of psychological therapy evaluation.

Until now, psychologists use mood evaluation scales and psychiatric diagnoses [13] to assess personality or mood. However, multiple sources of potential error are associated with these techniques, particularly in relation to the ability and willingness of persons to communicate about their actual state of mind [14].

A more reliable method is the degree of subjective arousal experienced by a client. This can be measured by means of a clinical self-evaluation test: the *Subjective Unit of Distress (SUD)* [15]. It consists of a score form on which the participant can mark his or her level of experienced tension, on a scale of 1 to 10.

2. Emotional Prosody Measurement (EPM)

We will now discuss how to do EPM in general. In addition, we discuss the mainly automated procedure to follow for such a psychological therapy evaluation method.

A sample of speech has to be (preferably digitally) recorded. If present, speckle noise and other voices have to be removed. Optionally, a Low Pass Filter (LPF) can be applied. A Fourier Transformation is applied to extract the F0-signal from the (complete) speech signal. Last, the variability of the F0 should be determined by, for example, the *standard deviation (SD)* of the F0-signal (*SD F0*). See also Figure 1 for the EPM processing scheme.

For therapy-evaluation at least two speech samples are needed. For example, during the take-in or during the first therapy-session (at the start of the treatment) a first recording could be made. A second recording could, for example, be made during the tenth session, when progress is expected or when therapy is ended. The EPMs of both recordings should be compared using statistics to determine whether a difference is present in the F0-signal. Preferably, the subject of speech should be the same; for example, the problems that initiated the treatment.

Figure 1. Processing scheme for Emotional Prosody Measurement (EPM).

[1] Source: Merriam-Webster Online Dictionary, URL: *http://www.m-w.com* .

3. Applying EPM on Mimicked Emotions

So far, the use of EPM in therapy evaluation is only made plausible and has not been proven. In multiple studies done in the past [12] actors were asked to induce emotions. In our study, however, the six participants (both men and women; age 20-50) were not skilled in expressing emotions by voice. They were asked to read aloud a story in a sad and in a happy manner. The difference between both conditions was as clear as it was consistent between the participants (e.g., see Figure 2).

These findings are in line with the existing theoretical distinction of emotional and emotive communication [16]. Emotional communication is a type of spontaneous, unintentional leakage or bursting out of emotion in speech, while emotive communication has no automatic or necessary relation to 'real' inner affective states, but is a strategic signalling of affective information in speaking [17]. Mimicking of emotions triggers emotive communication, where our interest is emotional communication.

So, this approach was in our opinion not satisfying since it was far from being ecological valid. Mimicking sad and happy speech should be considered as emotive communication. Hence, it does not resemble sad and happy speech as it is in real life situations (i.e., emotional communication). Therefore, an experimentally controlled and ecological valid research method was developed.

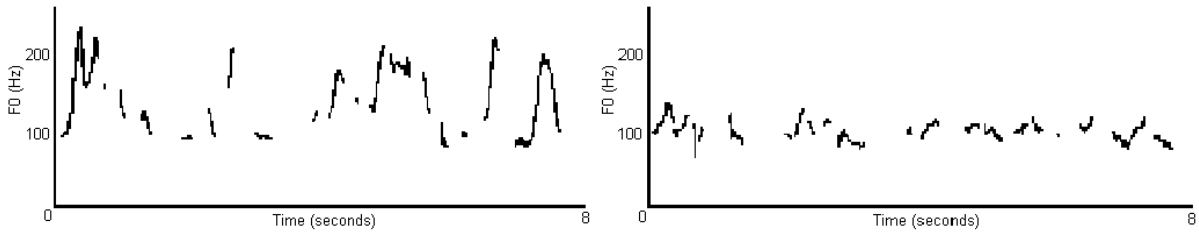

Figure 2. The raw F0-signals of respectively a mimicked happy (left) and mimicked sad (right) voice.

4. Method of Validation of EPM for Therapy Evaluation

4.1. Design and Procedure

Despite appealing results in the pilot study, another method for the validation of EPM for psychological therapy evaluation needed to be designed. Two methods emerged: 1) Telling stories, which do invoke true emotions, resulting in emotional communication and 2) Triggering real emotions by reliving of experiences.

A trade-off between both methods was hard to make, since the first could be controlled very nicely and the second would be an ultimate attempt in doing ecological valid research [12]. Therefore, we have chosen to include both methods as two more or less separate blocks within one experiment. Both before and after these two blocks a neutral story should be read as baselines. The baselines, a pre-test and a post-test condition, were used to measure changes in participants' behaviour during the intervening time of the main experiment. Each of the blocks consists of an anxiety triggering condition and of a neutral or even happy condition.

The fictive stories used in the main experiment, that had to be read aloud by the participants, were controlled both on both syntactic structure and on the complexity of the

line of story. The words used were also controlled on the frequency of appearance in everyday use and on complexity of pronunciation. Additionally, the stories were controlled for being respectively anxiety triggering (by use of anxiety triggering words) and being neutral. For the full stories we refer to [18].

The anxiety triggering conditions should be compared with a take-in or a first therapy session in which the trauma is discussed. The neutral or even happy conditions should be comparable with a moment in therapy when the trauma is (partly) relieved. This made sense because the anxiety and the accompanying stress should be declined when a trauma is (partly) relieved.

To control for effects due to the order of the conditions, the two condition blocks, as well as the two conditions within them, were counterbalanced across clients. The experienced stress was measured by the SUD. For each minute the subjects were asked to grade the degree of experienced distress on a scale from 1 to 10. During the complete experiment one continuous recording was made. However, to make a one-on-one mapping between EPM and SUD possible, the phases of recording during the four core conditions of the experiment were divided in samples of one minute of speech each, making up a total of twelve speech samples. The complete experiment, including practice session and baselines, took about 40 minutes. After completion the result was a unique database of speech samples, extremely useful for EPM [19].

4.2. Participants and Apparatus

Twenty-five Dutch females (average age of 38) participated. All subjects were diagnosed as having a Panic Disorder with Agoraphobia. They suffer of recurrent and unexpected Panic Attacks. These are accompanied by anxiety about, or avoidance of, places or situations from which escape might be difficult or in which help may not be available in the event of having a Panic Attack or panic like symptoms [20]. This group of clients was relatively sensitive to stress inducing factors and did have a low threshold for becoming stressed. This made this group an ideal group of participants for the reliving block, In addition, they were also expected to react on the anxiety-triggering words embedded in the anxiety-triggering story.

All participants took part in the experiment on a voluntary basis, with consent for the audio recording of the experimental session and the use of these recordings for scientific research. The specific goal of the experiments was revealed neither before nor after the sessions. However, a more general aim of the study, i.e. the improvement of therapy evaluation techniques, was indicated in advance.

The recording equipment consisted of a personal computer, a microphone pre-amplifier, and a microphone. The recording was done with a sample rate of 44.1 kHz, mono channel, with a resolution of 16 bits. The recordings were divided in files of approximately one minute of speech (for more info see [18]).

5. Results

5.1. Validation of the Experimental Manipulation of Stress

We analyzed the SUD data by means of a Repeated Measures Multivariate ANOVA to determine if the manipulation of stress has been successful. A strong difference in SUD was found between the two blocks ($F(1,23)=8.176$, $p<.009$). In addition, a main effect between the conditions within the blocks was found, which indicated that our experimental

manipulation of stress in the experiment was effective ($F(3,21)=14.908$, $p<.001$). On average, the SUD in the stress triggering conditions (3.76) was higher than in the neutral conditions (2.83). The SUD generally increased over time in the stress triggering conditions, while it decreased in the neutral conditions. Hence, the database of speech samples was validated for research towards EPM as a therapy evaluation method.

5.2. Validation of EPM as a Therapy Evaluation Method

The SD F0 data was analyzed by means of a Repeated Measures Multivariate ANOVA to determine its behaviour compared to the SUD. In the story telling block, the average SD F0 of the stress triggering condition (36.93) was lower than that of the neutral condition (39.78) ($F(1,20)=13.229$, $p<.002$), while no difference was found between the two conditions in the reliving block.

An additional Repeated Measures Multivariate ANOVA, both including Condition and Time as within participant factors, was performed for both blocks. The average values of both measures in the story telling block showed an opposite relation between both conditions ($r= -0.82$, $p<.022$, $n=6$).

The SUD was lower in the neutral condition (2.70) than in the stress triggering condition (3.52) ($F(1,23)=4.417$, $p<.047$), whereas for the SD F0 an opposite effect was found. These findings were, however, not clearly present in the reliving block. An additional analysis resolved that the order of conditions did have a strong influence on the measured SD F0 ($F(21,32)=4.021$, $p<.001$).

In a subsequent analysis we, therefore, analyzed the four conditions using their order of appearance for each participant as a criterion. With this the difference in experimental manipulation between both the blocks and the conditions within them was more or less ignored. So, the focus of interest lies solely on the comparison of both measures during and between the conditions over time.

During all four core conditions a sharp increase in SUD was present ($F(1,23)=9.953$, $p<.001$). In three of the four core conditions, the SD F0 showed a sharp and almost linear decrease ($F(1,23)=5.618$, $p<.007$). In addition, a clear and stable increase in SD F0 was present during all five breaks between the sequential conditions ($F(1,23)=9.138$, $p<.006$).

A significant correlation of -0.59 ($p<.023$, $n=12$) was found between SD F0 and SUD over all four conditions. An even stronger correlation arose between SUD and SD F0 of -0.88 ($p<.011$, $n=6$) for the first block, while a correlation of -0.69 ($p<.064$, $n=6$) was found for the second block (see Figure 3).

6. Conclusions

The experimental manipulation of anxiety/stress has been very successful, which resulted in a database with speech samples suitable for doing research in EPM. However, stronger induced effects on the SUD were present in the reliving block than in the story telling block. This is in accordance with previous findings [21,22] that emotions experienced in more realistic conditions are much stronger than those in experimental conditions.

The comparison of both SUD and EPM revealed the expected behaviour in the story telling block. In the reliving block, however, initially no difference was found in the SD F0 between the stress triggering condition and the neutral condition, which contradicted the data of the SUD. However, the difference in measured tension by the EPM and the SUD in the reliving block can be explained by referring to the intensity of the emotions triggered

during the two conditions. In both stress triggering conditions of both blocks the increasing degree of tension can be ascribed to the stress triggered by anxiety; in the neutral reliving condition, on the other hand, an intense reliving of positive emotions was present that also resulted in a high degree of (positive) tension. According to the EPM, only in the neutral story telling condition a relief in tension was present.

So, we may conclude that the EPM did measure the amount of tension, but does not discriminate tension as a consequence of pleasant or non-pleasant emotions. The SUD on the other hand, does react on differences in the source of tension.

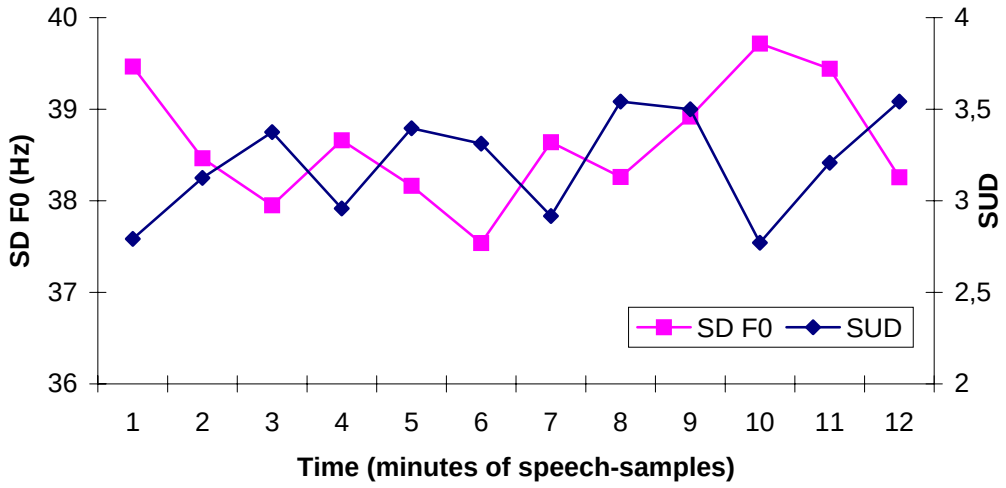

Figure 3. The average normalized Subjective Unit of Distress (SUD) and the Emotional Prosody Measurement (EPM) by the standard deviation of the F0 (SD F0) of 12 minutes of 'clean' speech-samples during the experiment.

An additional post-hoc test showed that the order of presentation of the conditions had an effect on the EPM. We, therefore, checked how the degree of tension experienced by the participants varied across the session not just across blocks but from minute to minute, using both the SUD and EPM. A strong effect of time on the observed degree of stress, as indicated by both measures, was found. Whereas the degree of stress measured by SUD and EPM slowly declined across the experiment as a whole, the stress experienced by the participants, according to both measures, increased time and again during the different conditions of the experiment and decreased in the time between the conditions (see Figure 3).

With this analysis the opposite behaviour of both SUD and the variability of F0 became even evident. However, this pattern of changes in the EPM was somewhat less clear in the second block. Possibly, the long duration of the experiment gradually resulted in less concentration, leading to a decline in the intensity of experienced emotions in the second half of the experiment. The analysis of SUD and EPM across time validated the EPM as an indicator for the presence of anxiety/stress and thus confirmed the importance of research with respect to (indirect) physiological measures, such as the EPM, for the diagnosis of people's emotional and mental state.

7. Discussion

The incorporation of direct physiological measures (EMG, EEG, etc.) is not easy from practical, technical, and financial points of view and clients feel uncomfortable when they are connected to all kinds of machinery. Audio recordings and their automated analysis, on the other hand, do not lead to discomfort and are suitable in a wide range of settings. They require only a small endeavour of the therapist and provide him a powerful additional diagnostic tool.

Other potential benefits of EPM, lie in the determination of the coping style of clients and in the determination of the severity of clinical disorders. Additionally, the EPM can probably make a differentiation between emotive and emotional communication.

Note that the main strength of the EPM lies not in making the distinction between emotional states, but in measuring the intensity of emotions. This is conform previous findings, which state that the variability of F0 mirrors the amount of anxiety/stress present [12]. The main strength of the SUD (Subjective Unit of Distress), on the other hand, was that it makes the distinction between levels of experienced (dis)comfort with the emotions. In combination, the two measures can, therefore, make a complete and reliable diagnosis of both the coping style and the emotional well being of persons.

Another perspective is that of 'normal' communication. People have the ability to take into account the emotional state of their partner in conversation, which makes it possible to adapt their communication style. Integration of EPM with speech recognition techniques could advance human-computer interaction by way of speech [19,23,24,25,26].

So, we can conclude by stating that a unique experimental setup was presented that resulted in a database of both experimentally controlled and ecological valid speech samples. This was of the utmost importance, since research depends on it [19]. It was used to prove that EPM (preferably combined with SUD) is an objective, easily usable (e.g., compared to [27]), and enormously powerful evaluation method for a wide range of settings [19,23,24,25,26], such as psychological therapy [28].

8. Acknowledgements

We would like to thank the Angstpolikliniek GGZN for their approval on using the addresses of their clients. In addition, we would like to thank the Muriel Hagenaars who guided the experiments. Ton Dijkstra is gratefully acknowledged for a range of efforts during the research project Trauma and Language 1. Last, we would like to thank all others who made the project Trauma and Language 1 possible.

9. References

[1] Helmholtz, H. von (1896). *Die Lerne von den Tonempfindungen als physiologische Grundlage fur die Theorie der Musik.* (5th ed.). Braunschweig: Friedrich Vieweg und Sohn. (Original work published in 1863.)

[2] Stevens, K.N. and House, A.S. (1955). Development of a quantitative description of vowel articulation. *Journal of the Acoustic Society of America, 27*, 484-493.

[3] Chiba, T. and Kajiyama, M. (1941). *The vowel: Its nature and structure*. Tokyo, Japan: Tokyo-Kaiseikan Publishing Company, Ltd.

[4] Fant, G. (1960). *Acoustic theory of speech production*. The Hague, The Netherlands: Mounton and Co.

[5] Brunswik, E. (1956). *Perception and the representative design of psychological experiments.* Berkeley: University of California Press.

[6] Gifford, R. (1994). A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behaviour. *Journal of Personality and Social Psychology*, *66*, 398-412.

[7] Hammond, K.R. & Stewart, T. R. (2001). *The essential Brunswik: Beginnings, explanations, applications*. New York: Oxford University Press.

[8] Berlin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*, 129-135.

[9] Ellis, A.W. (1989). Neuro-cognitive processing of faces and voices. In H.D. Young and A.W. Ellis (Eds.), *Handbook of Research on Face Processing* (pp. 207-215). Amsterdam, The Netherlands: Elsevier Science Publications B.V.

[10] Fitch, W.T. (2000). The evolution of speech: A comprehensive review. *Trends in Cognitive Sciences*, *4*, 259-267.

[11] Monrad-Krohn, G.H. (1963). The third element of speech: Prosody and its disorders. In L. Halpern (Ed.), *Problems of Dynamic Neurology* (pp. 101-117). Hebrew, Israel: Hebrew University Press.

[12] Scherer, K.R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*, 227-256.

[13] Knapp, S., & VandeCreek, L. (1994). *Anxiety disorders: A scientific approach for selecting the most effective treatment.* Sarasota, Florida, U.S.A.: Professional Resource Press / Professional Resource Exchange, Inc.

[14] Sackheim, H.A., & Gur, R.C. (1978). Self deception, self confrontation, and consciousness. In G.E. Schwartz & D. Shapiro (Eds.), *Consciousness and self regulation: Advances in research* (Vol. 2, pp. 117-129). New York: Plenum Press.

[15] Wolpe, J. (1958). *Psychotherapy by reciprocal inhibition.* Stanford, California, U.S.A.: Stanford University Press.

[16] Marty, A. (1908). *Untersuchungen zur allgemeinen grundlegung der grammatik und sprachphilosophie*. Halle/Saale, Germany: Niemeyer.

[17] Caffi, C., & Janney, R.W. (1994). Toward a pragmatics of emotive communication. *Journal of Pragmatics*, *22*, 325-373.

[18] Broek, E.L. van den (2001). *Diagnosis of emotional state using pitch analysis and the SUD: An exploratory feasibility study*. M.Sc.-thesis Faculty of Social Sciences, University of Nijmegen, Nijmegen, The Netherlands.

[19] Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Toward a new generation of databases. *Speech Communication*, *40*, 33-60.

[20] American Psychiatric Association (1996). *DSM IV Sourcebook: Volume 2* (1st ed., rev.). Washington, DC, U.S.A.: American Psychiatric Association.

[21] Scherer, K.R. (1981). Vocal indicators of stress. In J.K. Darby (Ed.), *Speech evaluation in psychiatry* (pp. 171-187). New York: Grune & Stratton, Inc.

[22] Streeter, L.A., Macdonald, N.H., Apple, W., Krauss, R.M., & Galotti, K.M. (1983). Acoustic and perceptual indicators of emotional stress. *Journal of the Acoustical Society of America*, *73*, 1354-1360.

[23] Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication*, *36*, 31-43.

[24] M. Schröder & J. Trouvain (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6, pp. 365-377.

[25] Yacoub, S., Simske, S., Lin, X., & Burns, J. (2003). *Recognition of emotions in interactive voice response systems*. Hewlett-Packard Development Company, L.P. Report number HPL-2003-136.

[26] HUMAINE: The European Network of Excellence on emotion research and human-machine interaction, URL: http://emotion-research.net/ [last accessed on 22 March 2004].

[27] Bostanov, V. & Kotchoubey B. (2004). Recognition of affective prosody: Continuous wavelet measures of event-related brain potentials to emotional exclamations. *Psychophysiology*, *41*, 259-268.

[28] Kucharska-Pietura K., Nikolaou V., Masiak M., & Treasure J. (2004). The recognition of emotion in the faces and voice of anorexia nervosa. *International Journal of Eating Disorders*, *35*, 42-47.